A Hybrid Metaheuristics based technique for Mutation Based Disease Classification

Original Scientific Paper

Manu Phogat
Guru Jambheshwar University of Science & Technology,
Hisar, India-125001
kunjean4181@gmail.com

Dharmender Kumar
Guru Jambheshwar University of Science & Technology,
Hisar, India-125001
dharmindia24@gmail.com

Abstract *– Due to recent advancements in computational biology, DNA microarray technology has evolved as a useful tool in the detection of mutation among various complex diseases like cancer. The availability of thousands of microarray datasets makes this field an active area of research. Early cancer detection can reduce the mortality rate and the treatment cost. Cancer classification is a process to provide a detailed overview of the disease microenvironment for better diagnosis. However, the gene microarray datasets suffer from a curse of dimensionality problems also the classification models are prone to be overfitted due to small sample size and large feature space. To address these issues, the authors have proposed an Improved Binary Competitive Swarm Optimization Whale Optimization Algorithm (IBCSOWOA) for cancer classification, in which IBCSO has been employed to reduce the informative gene subset originated from using minimum redundancy maximum relevance (mRMR) as filter method. The IBCSOWOA technique has been tested on an artificial neural network (ANN) model and the whale optimization algorithm (WOA) is used for parameter tuning of the model. The performance of the proposed IBCSOWOA is tested on six different mutation-based microarray datasets and compared with existing disease prediction methods. The experimental results indicate the superiority of the proposed technique over the existing nature-inspired methods in terms of optimal feature subset, classification accuracy, and convergence rate. The proposed technique has illustrated above 98% accuracy in all six datasets with the highest accuracy of 99.45% in the Lung cancer dataset.*

Keywords*: Feature Selection, Metaheuristic, Competitive Swarm Optimization, Whale Optimization Algorithm, Artificial Neural Network, Hybrid Techniques, Classification*

1. INTRODUCTION

Mutations are types of abnormal changes in the genetic material, especially within nucleic acid (RNA, DNA). The change in the nucleotide sequence of DNA results in an alteration in the amino acid's sequences of proteins, which leads to certain genotypic and phenotypic changes in the human body. Various diseases are caused by mutation, the most common are different types of cancers. Cancer is caused by acquired mutations, also known as somatic mutation, that occurs due to changes in nucleotide patterns such as C -> T [1]. The tumor is a very complex disease, which is driven by various factors like lifestyle, environment, and genetics. During the past few years, investigation on gene mutation at both specific loci and large scale has been carried out to increase the knowledge of molecular diversity in complex diseases like cancer. Several large-scale cancer genome projects have been carried out which have provided huge amounts of high-dimensional data. These projects include ICGC (The international cancer genome Consortium), TGCA (The Cancer Genome Atlas), and Cancer Genome Project by Trust Sanger Institute along with many other experiment-based studies that have been conducted for cancer classification [2-4]. The tremendous rise in DNA Microarray technology has given us a deep insight into various alterations and genetic variations that helps us in the early detection of complex diseases such as cancer. The genomic datasets have thousands of genes and a small number of samples, which makes the selection and classification of cancer very difficult. The high-dimensional datasets contain redundant and irrelevant genes that decrease the training strength of a classifier. So numerous techniques have been proposed in recent years for cancer classification using machine learning [5]. The researchers found machine learning, the most significant tool to perform data analysis in

biological datasets [6]. The main objective of cancer classification is to identify the biomarkers (genes) to differentiate various types of cancer. Techniques that have been proposed in the literature for gene selection methods are categorized into four methods that are as follows: wrapper, filter, embedded, and hybrid. The filter method selects the features on the basis that how to correlate with the output or based on their relationship with the output. The filter methods are fast, classifier-independent, and computationally inexpensive [7]. On the contrary wrapper methods split the data into subsets and train a model using this, the addition and deletion of features are dependent on the output of the model. So generally, wrapper methods provide better accuracy than the filter methods, but in comparison, they are computationally expensive because they tested all possible combinations of feature subsets. The hybrid methods combine the qualities of both filter and wrapper methods to create the best sub-set of features. In hybrid methods, the researchers first apply the filter method to reduce the feature size and then use the wrapper technique for the final selection of the relevant genes. The carcinoma microarray datasets are highly dimensional and suffer from a curse of dimensionality problem, which means a high number of features and a small sample size. The small p and large n state the cancer classification problem as an NP-hard problem. For the past few years, different metaheuristics techniques have been proposed in the literature to solve the different variety of problems related to real-life. The metaheuristics algorithms are easily understandable and computationally inexpensive because they provide an optimal solution in a decent amount of time, which makes them pretty useful in the problem areas of bioinformatics and computational biology. So, for the selection and classification of genes metaheuristics algorithms are used in the filter, wrapper, and hybrid techniques, and also used for parameter tuning of a classifier to improve the classification accuracy. The wrapper techniques generally used the conventional fitness function for selecting the genes, to maximize the performance of a classifier. Accordingly in this study, the authors have introduced a new fitness function to overcome the limitation of conventional fitness or basic fitness function. In the past few years, machine-learning algorithms have been used for optimal prediction in the field of computational biology. It is difficult to identify the gene signification for complex and large biological structures, so the metaheuristics techniques add new insight. Numerous nature-inspired techniques are used in cancer classification for relevant gene selection [8].

In the literature part authors have introduced a dominant metaheuristics algorithm, CSO for solving real-world problems, also due to its impressive capabilities it has also been used for gaining the nearest optimal solution in the large search space efficiently. The CSO Technique is highly effective to find the relevant gene subset, but this process needs a few improvements for slow convergence, which inspired us to search for further progress. Another popular metaheuristic technique is the whale optimization algorithm (WOA) widely used for hyperparameter tuning in various classifiers [9]. WOA leads to global optima and demonstrates rapid convergence as compared to other metaheuristic techniques. Moreover, WOA is very flexible and robust in handling a large number of decision variables.

In recent scenarios, various wrapper techniques are introduced in bioinformatics to examine the biological system for a thorough perspective. Many benchmark wrapper techniques have been explored for mutated gene selection in the tumor classification but these algorithms failed to identify the correlation between genes in their search process [10]. This leads to a rise in the computational load for identifying the optimal genes. To overcome this weakness, researchers have been exploring various hybrid evolutionary methods, such as a hybrid of PSO & GA, a hybrid of BBHA & BPSO, a hybrid of TLBO and GSA and fusion of EFS and AGOA Algorithm [11-14].

The existing hybrid methods still experience a lot of shortcomings, such as being stuck in local optima and having high execution time, so they do not accomplish adequate classification accuracy. In virtue of that, our study developed a new hybrid metaheuristics technique to conquer the shortcomings of the traditional algorithms and identify the target genes for accurate cancer prediction. The proposed technique has some impeccable advantages such as scaling down the computational complexity and boosting the accountability of the dataset; also, in addition, it can handle high dimensional data with the optimal solutions in a feasible time.

The key input of the present study is as follows:

- This article introduced a new hybrid metaheuristics technique with the combination of IMCSO and WO algorithms.
- The proposed technique introduces a new fitness function to improve classification performance
- The proposed technique optimizes the hyperparameter of the ANN classifier.

The rest of the arrangement of this article is as follows: Section 2 Comprises Related work in this the authors give a brief introduction of IBCSO, WO, and ANN techniques followed by a proposed hybrid metaheuristics technique (IBCSOWO) with an ANN model and its advantages. Section 3, focused on experimental results and discussion. Section 4 elaborates on the conclusion and future scope.

2. RELATED WORK

In the past few years, machine-learning algorithms have been used for optimal prediction in the field of computational biology. It is difficult to identify the gene signification for complex and large biological structures,

so the metaheuristics techniques add new insight. Numerous nature-inspired techniques are used in cancer classification for relevant gene selection. A hybrid gene identification technique has been proposed by, using the fusion of an Artificial bee colony (ABC) and a Genetic Algorithm (GA) [15]. The main goal of the article was to combine the advantages of both techniques to predict the relevant genes. Another gene selection technique proposed by using PSO and KNN target genes subset for tumor classification can be identified [16]. Another hybrid method using IGWO and PSO algorithms for the prediction of the most relevant genes in breast tumor classification can be acknowledged [17]. In another hybrid metaheuristics technique, Sharma et al. proposed a multi-objective framework (C-HMOSHSSA) with the fusion of unique MOSHO and SSA (Salp Swarm Algorithm), to select the optimal set of genes from high dimensional datasets [18].

Table 1. Comprises of Comparison of various tumor classification Techniques.

References	Techniques	Advantages	Disadvantages
[19]	Stacked Auto-Encoder (Deep Learning)	Formulate a clinical decision support system (DSS) to aid pharmacologists.	The performance of the model with respect to training time is poor than the existing models
[20]	Cuckoo search with crossover (Classification)	Assist both microarray and NGS-based miRNA expression data.	No Sensitivity analysis of feature subset.
[21]	Binary Bat Algorithm with SVM	Greedy Crossover proposed to rearrange the sub-optimal solutions. Overcome premature convergence	The technique is dependent on a particular dataset.
[22]	mRMR with Modified BAT Algorithm (Classification)	DNA microarray appearances empowered the simultaneous observation of expression levels of a large number of genes.	The time complexity of the proposed technique is high.
[16]	PSO and BAT (Classification)	The heuristics search technique is used to select the optimal values of K.	Cross-validation not performed, Higher complexity
[23]	Feature Score and ACO	The subset genes sampled are mapped into a dissimilarity space. Classifiers surpass the feature-based models.	The proposed technique only used one filter and wrapper technique in the study.
[24]	IG and GA (Classification)	Feed Forward neural network is used that gives good classification accuracy.	Smaller sample size with high complexity.
[25]	Gene Bank and GSA	The adaptive distance technique is used to improve the performance of the algorithm.	Slow convergence and high complexity.

The recent literature depicts the latest machine learning techniques and metaheuristics methods, which are applied, for cancer classification and shows potential results but most of them suffer from the local optima stagnation, redundancy, and slow convergence rate as depicted in Table 1. The feature space of microarray cancer datasets is large so the author used a hybrid feature selection technique to strike out the optimal feature subset.

Binary Competitive Swarm Optimization (BCSO)

Competitive Swarm Optimization is proposed by Cheng and Jin in 2015, it is a popular algorithm that uses a pair-wise competitive scenario. The CSO is considered a novel version of Particle Swarm Optimization [26]. The CSO randomly divides the population of particles into two equal-size groups. Each group is in a competitive spirit with each other and out of this competition, a particle having better fitness value is considered as a winner and directly moves to the next level. The loser particle updates its velocity and position by attaining information from the winner.

The loser velocity is updated as:

$$v_{l,d}(i+1) = r_1 v_{l,d}(i) + r_2\left(x_{w,d}(i) - x_{l,d}(i)\right) + \alpha r_3\left(\bar{x}_d(i) - x_{l,d}(i)\right) \quad (1)$$

$$x_{l,d}(i+1) = x_{l,d}(i) + v_{l,d}(i+1) \quad (2)$$

Where v_l and x_l are the velocity and position of the loser particle, x_w is the position of the winner particle, $\bar{x}$ is the mean position of the current swarm, r_1, r_2, and r_3 are three independent random vectors distributed in [0,1], α is the social factor, d is the dimension of search space and i is the iteration number.

The conventional CSO converts into binary CSO when the continuous real domain is converted into a discrete domain, so the solution can be represented in binary form. Traditionally the wrapper techniques consider Binary CSO instead of CSO. The solution represented in BCSO is either 0 or 1.

The main steps of BPSO are as follows:

- At first, randomly initialize the population of N particles. The velocity of every particle will be considered zero V=0.
- For every particle fitness is evaluated; the best fitness score particle is named gbest.
- Divide the particles into two groups on each iteration.
- The velocity of the loser particle is updated using Equation 1.
- Now the velocity is transformed into a probability value between [0,1], using a transfer function.

So, the updated position is calculated as:

$$x_{l,d}(i+1) = \begin{cases} 1, & (if\ S(v_{l,d}(i+1) > r_4) \\ 0, & (otherwise) \end{cases} \quad (3)$$

Here S is the transfer function and r_4 is a random vector dispersed in [0,1].

The BCSO is applied for the feature selection process in classification tasks. In BCSO the bit value 1 indicate feature is selected, while bit 0 is for the unselected feature [26].

Whale Optimization Algorithm (WOA)

The whale optimization algorithm introduced by Mirjalili and Lewis in 2016, is based on the social behavior of humpback whales [27]. The WOA is robust and easy to implement when compared to other nature-inspired techniques. When a whale attacks their prey, they encircle it and swim up to the surface in a shrinking circle. The WOA works in three phases: a) Shrinking encircling prey, b) Spiral shape attacking method, and c) search for prey.

The encircling behavior is expressed as:

$$\vec{D} = \left|\vec{C}\,\vec{X^*}(t) - \vec{X}(t)\right| \quad (4)$$

$$\vec{X}(t+1) = \vec{X^*}(t) - \vec{A}.\vec{D} \quad (5)$$

$$\vec{A} = 2\vec{a}.\vec{r} - \vec{a} \quad (6)$$

$$\vec{C} = 2.\vec{r} \quad (7)$$

Where $\vec{X}$ represents whale position, $\vec{X^*}$ is the general best position, a represents linearly reduced distribution within [2, 0] during iterations, t stands for present iteration, and r is random not dispersed within [0,1].

The helix shape movement of the whales inspired the spiral model, the Spiral Shape attacking method (exploita-tion phase) equation as follows:

$$\vec{X}(t+1) = \vec{D'}.e^{bl}.\cos(2\pi l) + \vec{X^*}(t) \quad (8)$$

Where $\vec{D'}$ is a vector that stores the absolute distance between $\vec{X^*}(t)$ and $\vec{X}(t)$, b is a fixed value that interprets logarithmic spiral space and l is a random numerical value ranged between [-1, 1].

If A>1 or A<-1, so for global optimizers, a search agent is revised as per a random search agent in place of the best search agent. The search for prey model equation is as follows:

$$\vec{D} = \left|\vec{C}.\overrightarrow{X_{rand}}(t) - \vec{X}\right| \quad (9)$$

$$\vec{X}(t+1) = \overrightarrow{X_{rand}}(t) - \vec{X}.\vec{D} \quad (10)$$

The $\overrightarrow{X_{rand}}$ is called promptly from whales in the current iteration.

Artificial Neural Network

The ANN model proposed by McCulloch and Pitts in 1943, is based on the functionality of biological neurons [28]. The neural network consists of input neurons, which consist of input and their weight. Next is the internal neuron that provides a function, which has the summation of all weights and biases. The last is the output neuron in which the summation of weights and biases are passed through an activation function Figure 1.

The products of ANN with K elements are given as follows:

$$y(x) = \sum_{i=1}^{k} wiyi(x) \quad (11)$$

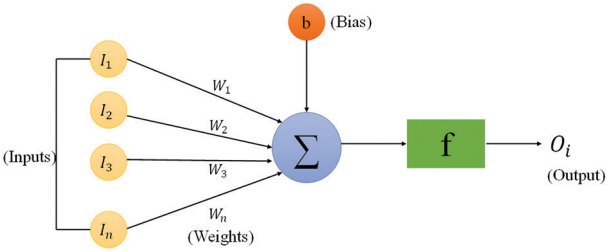

Figure 1. General Structure of ANN Network

The ANN has various architecture models such as single-layer feed-forward (perceptron), feed-forward neural network, and recurrent neural network. Several different variants are there of these architectures such as Kohonen networks, convolution neural networks, extreme learning machines, Hopfield networks, etc. In this article, a Multilayer Perceptron model is used which is part of a feed-forward neural network, the Multi-Layer Perceptron consists of an input layer, a hidden layer, and an output layer, and each node is associated with an activation function. The ANN has been widely used for tumor classification and the other different fields of bioinformatics [29].

3. PROPOSED FRAMEWORK

The proposed technique has three main aspects such as preprocessing, optimization, and classification. In the preprocessing phase, the input datasets go through a filter technique mRMR (minimum Redundancy Maximum Relevance) for gene selection. The genes are categorized according to their rank by filter method, followed by which a reduced dataset is obtained.

Pre-Processing Phase

The filter-based feature selection techniques are highly effective for filtering out the irrelevant and redundant genes in tumor classification. The efficient mRMR technique is used to generate the subset of high-quality genes. In mRMR the mutual information of variables X and c is determined based on the entropy of both X and c for each gene variable [30]. The entropy equation is followed as:

$$I(c;X) = H(c) - H\left(\frac{c}{X}\right) \quad (12)$$

Where $H(c)$ is entropy and $H(\frac{c}{X})$ is conditional entropy between variables and class. The main concept of minimum redundancy is to find out the genes which are

mutually maximally different from others. The average minimum redundancy is given as:

$$minZ(X,c) = \frac{1}{|s^2|}\sum_{x_j \in s} I\,(x_j ; x_k) \quad (13)$$

Where s stands for a subset of required genes and(x_j, x_k) represents mutual information between j-th and k-th genes. Again, the concept of mutual information is needed to choose a subset S with N genes having a maximum dependency on target class c. The average maximum relevance is formulated as:

$$maxV(X,c) = \frac{1}{|s|}\sum_{x_j \in s} I(\,c, x_j) \quad (14)$$

These two conditions are combined into a single criteria function $Max(V\text{-}Z)$, here mRMR is for discrete variable form. The equation for mRMR is an integration of Equations 13 and 14 and is described as:

$$j_{mRMR}(\phi) = I(c;X)^{-1}/_{|S^2|}\sum_{x_j \in s} I(x_j; x_k) \quad (15)$$

x_j is a selected subset of genes Sand x_k is the original genes set.

Proposed Algorithm

The mRMR filter method effectively reduces the dimension of the dataset and hands over the important genes. Now the effective classification model is applied to further scaling down the dimension of the gene subset and a hybrid algorithm is used with an ANN model to accomplish the maximum classification accuracy.

A proposed hybrid model called IBCSOWOA is used by fusion of two metaheuristics algorithms namely IBCSO and WOA, the IBCSO is used as a wrapper technique to deal with the dimension reduction and WOA is used for parameter tuning to improve the classification accuracy of the model. Algorithm 1 can show the Pseudo-code of the proposed algorithm. The IBCSO (Improved Binary Competitive Swarm Optimization) is an enhanced version of BCSO in which the swarm population is divided into three different swarms to escalate the convergence speed. The idea of tri-swarm in the proposed technique will allow two-thirds of the population to update, so more swarms will have a chance to move towards a good solution and it also balances exploration and exploitation. The mechanism of IBCSO describes in detail with the following steps:

- Initialization: At first the random particle is developed with a swarm size (m) in the multiple of three, which means the tri-division of the swarm. After initialization, all the particles are divided into three different groups.
- Tri-Competition: The particles from each group are picked up randomly and they go into a tri-competition. Out of the three only one will be declared the winner having the highest fitness value others will be named as the first and second losers.
- Updation: The winner is granted to pass immediately to the next iteration while both the loser is getting their velocity and position updated. The process of upgrading position and velocity is the same as in BPSO. The velocity and position for the first and second losers are described by the following equations: Loser1 ($l1$)

$$\upsilon_{l1,d}(i+1) = r_1(d,i)\upsilon_{l1,d}(i) + r_2\left(x_{w,d}(i) - x_{l1,d}(i)\right) + \alpha_1 r_3\left(\bar{x}_d(i) - x_{l1,d}(i)\right) \quad (16)$$

$$x_{l1,d}(i+1) = x_{l1,d}(i) + \upsilon_{l1,d}(i+1) \quad (17)$$

For Loser2 ($l2$)

$$\upsilon_{l2,d}(i+1) = r_4(d,i)\upsilon_{l2,d}(i) + r_5\left(x_{w,d}(i) - x_{l2,d}(i)\right) + \alpha_2 r_6\left(\bar{x}_d(i) - x_{l2,d}(i)\right) \quad (18)$$

$$x_{l2,d}(i+1) = x_{l2,d}(i) + \upsilon_{l2,d}(i+1) \quad (19)$$

Here $D(=m/3)$ represents the different swarms that participated in the competition. Position and velocities are $x_{w,d}(i)$, $x_{l1,d}(i)$, $x_{l2,d}(i)$ and $\upsilon_{w,d}(i)$, $\upsilon_{l1,d}(i)$, $\upsilon_{l2,d}(i)$ in the k-th round of competition (k=1,2,.....,k) in iteration i. r_1 to r_6 are six random numbers. α_1, α_2 are independent social factors regulating the impact of the mean position. $\bar{x}_d(i)$ is the mean position value of the relevant particle.

Algorithm 1: The pseudocode of Improved Binary Competitive Swarm Optimizer (IBCSO)

$P(i)$ represents the total swarm at each generation i. S represents a set of particles that do not participate in a Swarm. $X_w(i)$, $X_{l1}(i)$ and $X_{l2}(i)$ represent and two loser swarms respectively.

1. i=0;
2. Initialize the population p(0) randomly.
3. **While** termination criteria are not satisfied **do**
4. Search for fitness of every particle in $P(i)$;
5. $S=P(i)$, $P(i+1)=\phi$;
6. **While** $S=\phi$ **do**
7. Now arbitrarily choose three particles $X_1(i)$, $X_2(i)$ *and* $X_3(i)$ from S;
8. Arrange according to increasing order according to fitness function $f(X_1(i)) \le f(X_2(i)) \le f(X_3(i))$
9. Assign $X_w(i) = X_1(i)$, $X_{l1}(i) = X_2(i)$ *and* $X_{l2}(i) = X_3(i)$;
10. Add $X_w(i)$ *into* $P(i+1)$;
11. *Update* $X_{l1}(i)$, $X_{l2}(i)$ to $X_{l1}(i+1)$ and $X_{l2}(i+1)$ *by equation* (16)-(19) and add to $P(i+1)$;
12. Change the velocity into probability using the S-shaped threshold function
13. Upgrade the position of the loser using equation (3)
14. If the value of the position vector is 1 then the feature is selected, for 0 feature is unselected.
15. F_s= {*All Features where position vector*=1;}
16. Remove $X_1(i)$, $X_2(i)$ and $X_3(i)$ from S;
17. **End while**

18. i=i+1;

19. **End while**

20. **Return** F_s;

So, the main idea of IBCSO is to develop a subset of highly efficient gene selection techniques with a better convergence rate and a good balance between exploration and exploitation. For the classification tasks, ANN is used with Rectified linear unit activation function with WOA used for optimizing weights.

The equation for ReLU is as follows:

$$y = \max(0, x)$$

Where y is the output function and x is the input weight.

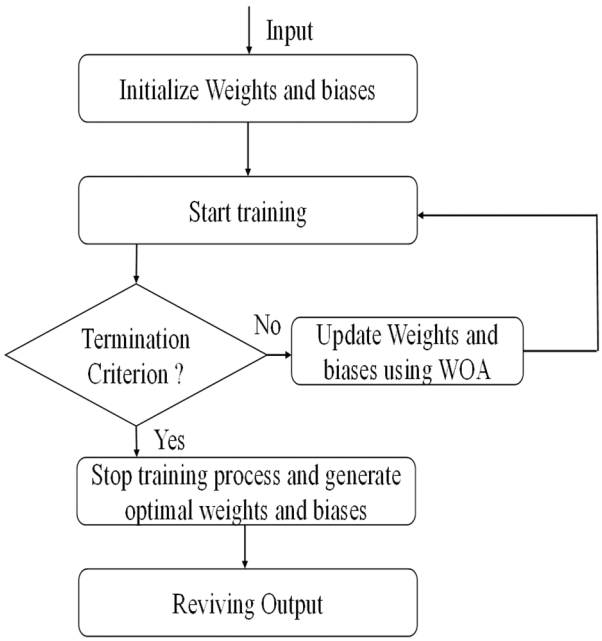

Figure 2. The Training Process of ANN

The training process of ANN is shown in Fig. 2. It takes input and achieves output based on current weights and biases. The computed output is compared with the target outcome with a loss function. After that, a backpropagation algorithm is used to update weight and bias in the next iteration. In the proposed technique WOA is used for the optimization of weights for the next iteration. The combined weights and biases in the ANN network are shown in Equation 21.

$$W = rs + 2r + 1 \tag{21}$$

Where s is the total no of input nodes and r is the total number of neurons in the hidden layer. The Mean Square Error (MSE) of ANN is the difference between predicted and actual values and it is used to alter the values of weights during backpropagation. MSE can be used by search agents (whales) as a fitness function to find out optimized weight values. The MSE can be calculated as:

$$MSE = \frac{\sum_{i=1}^{n}(O_i - O_i')^2}{n} \tag{22}$$

Where O_i is the actual output and O_i' is the predicted output for the input sample i. n is the number of samples.

Proposed fitness function

A new fitness function is proposed with respect to increasing the classification accuracy and reducing the number of genes. The fitness function is expressed as:

$$fitness(f) = \mathrm{A} * \frac{\mathrm{B}}{\Theta} + (1 - \mathrm{A}) * \Gamma \tag{23}$$

Here Γ represents the classification accuracy using ANN as a classifier. B Stands for the upper limit of selected features and A represent a constant value between 0 and 1. Θ is the measure of a chromosome length.

4. RESULT AND DISCUSSION

This section emphasizes the obtained results from the proposed technique and other methods on six mutation gene expression datasets from complex diseases such as Prostate, Lung, Breast, DLBCL, Arcene, and Dorothea. The Prostate dataset consists of 12600 genes with 136 samples out of which 77 include tumors and 56 are normal. The lung cancer dataset consists of 12533 genes with 181 tissue samples out of which 150 are of adenocarcinoma (AD-CA) and 31 are malignant pleural mesothelioma (MPM). The breast cancer dataset consists of 97 samples out of which 46 patients developed a tumor and the rest 51 are healthy samples, the number of genes in the dataset is 24481. The Diffuse Large B-cell lymphoma dataset (DLBCL) contains 11226 genes for 77 samples out of which 58 samples have large B cells and 19 are Follicular lymphoma. The Arcene dataset consists of 10000 genes and 100 samples, from which 56 are tumor samples and 44 are normal samples. Dorothea dataset consists of 800 samples and 100000 genes, from which 190 samples are positive and 610 are marked as negative. A detailed description of the datasets has been depicted in Table 2. The experimental results were obtained by using PYTHON 3.8, the observational evaluation was performed on NVIDIA Corporation TU104GL (Quadro TRX 5000) GPU and AMD 7662 (64-CORE*256) CPU along the Ubuntu 20.04.2. LTS (64-bit) operating system.

Table 2. Dataset description

No.	Datasets	Instances	Genes	Classes	Ref.
1	Prostate	136	12600	2	[31]
2	Lung	181	12533	2	[32]
3	Breast	97	24481	2	[31]
4	DLBCL	77	11226	2	[33]
5	Arcene	100	10000	2	[31]
6	Dorothea	800	100000	2	[31]

Parameter Setting

Table 3 contains the parameter values for IBCSO and WOA(ANN) techniques. The values were selected based on the results of considerable preliminary runs. The authors have tested the performance of all classifiers us-

ing 10-fold cross-validation to identify which one gives a better performance than the other methods on the top six-microarray datasets.

Table 3. Parameter Setting of the proposed technique.

S.no	Parameters	Value
1	No of generations	100
2	Population size (IBCSO, WOA)	200,100
3	Iterations	20
4	Θ Chromosome Length	50
5	Performance	Accuracy
6	$\alpha 1, \alpha 2$	0.2
7	Maximum velocity	6
8	α (WOA)	Linearly decreases from 2 to 0
9	b (WOA)	1

Evaluation Criteria

The performance of the proposed model is evaluated by the ANN classifier with the fitness function expressed in Equation 23. The proposed technique is evaluated by four different measures Specificity, Sensitivity, F-measure, Accuracy, and Matthews Correlation Coefficient (MCC). Performance measures are defined in the following equations:

$$Specificity(Se) = \frac{T_N}{T_N + F_P} \quad (24)$$

$$Sensitivity(Se) = \frac{T_P}{T_P + F_N} \quad (25)$$

$$F - score(F_s) = \frac{2*T_P}{2*T_P + F_P + F_N} \quad (26)$$

$$Accuracy(Acc) = \frac{T_P + T_N}{T_P + T_N + F_N + F_P} \quad (27)$$

$$MCC = \frac{(T_P * T_N - F_N * F_P)}{\sqrt{(T_P + F_P) + (T_P + F_N) + (T_N + F_P) + (T_N + F_N)}} \quad (28)$$

Here, T_P, T_N, F_P, and F_N are True positive, True negative, False positive, and False negative in all independent datasets. The T_p or T_N indicates exactly matched actual and predicted sample values, while F_p or F_N indicates distinct actual and predicted sample values.

Kappa Statistics

Cohen's kappa statistic is a performance model for classification, which measure the interrater reliability, which means a chance-corrected standardized measure of agreement between categorical scores produced by two raters. It measures value range from 0.0-1.0. The evaluation of Cohen's kappa is depicted as follows:

$$K = \frac{P0 - Pe}{1 - Pe} \quad (29)$$

Where $P0$ is the probability of agreement and Pe is the probability of random agreement.

Experimental Results and Analysis

The proposed technique is compared and validated with a series of algorithms frequently used on these datasets. The proposed technique is compared with different filter-based techniques such as FCBF, CMIM, mRMR, and Relief-F in Table 4. In the DNA microarray datasets, most of the genes are redundant, irrelevant, and noisy, so the top 100 genes are chosen to obtain the classification accuracy of all the methods. When IBCSOWOA is applied to the optimal gene set it increases the prediction accuracy and attains the highest accuracy 97.67% in the DLBCL dataset. Different types of nature-inspired algorithms are used in litera-ture for gene selection for complex disease classification.

In comparison with the earlier reported study, it should be found that using the mRMR filter selection method obtained 93.87% accuracy with ANN (WOA) on the DLBCL data; the proposed technique obtained the result of 97.67% accuracy with 100 selected genes on the same dataset. The highest performance is achieved by the proposed technique as 97.35% in Lung Cancer, 86.86% in Breast Cancer, 93.89% in Prostate Cancer, 94.25% in Arcene, and 91.45% in Dorothea dataset. This table states the efficiency results ANN (WOA) classifier except for the proposed technique in which a new fitness function is used; and chose the number of genes by CMIM, FCBF, mRMR, Relief-F, and the proposed technique of all datasets.

In Table 5 nature inspired algorithms such as PSO, GA, ACO, BCSO, and GWO are compared with the IBCSOWOA based on accuracy, standard deviation, and the optimal number of genes; the result demonstrates the effectiveness of the proposed technique. The optimum results between all gene selection techniques have been emphasized and marked in bold type. The most renowned metaheuristic algorithm such as GA selects the number of a gene to 18 in a DLBCL dataset with a classification performance of 97.52%. The classification performance of 99.45% in the Lung cancer dataset with 15 genes is the second maximum obtained by our proposed method. The metaheuristic algorithm PSO classification accuracy is 97.05% with 20 genes in the Lung cancer dataset.

In the ACO technique, the highest and lowest classification accuracy is 94.76% and 82.65% in DLBCL and Arcene datasets. Moreover, the BCSO algorithm obtained 96.21% classification efficiency from the DLBCL dataset with 22 genes and the GWO algorithm attain 91.25-classification accuracy from the Lung cancer dataset.

Table 4. Average classification performance with the top 100 genes from all six datasets

Techniques	DLBCL	Lung	Breast	Prostate	Arcene	Dorothea
FCBF	93.54	92.36	84.56	89.56	89.68	86.32
CMIM	92.68	92.99	83.75	87.34	90.25	85.65
mRMR	93.87	91.54	83.68	88.56	88.56	87.26
Relief-F	85.64	88.98	84.52	89.56	87.25	84.25
Proposed	**97.67**	**97.35**	**86.86**	**93.89**	**94.25**	**91.45**

Table 5. Observation of proposed technique with nature-inspired techniques with classification accuracy and STD, with the opti-mal number of selected genes.

Datasets	Performance metrics	PSO	GA	ACO	BCSO	GWO	Proposed
Prostate	#Acc ± STD # Genes	90.42± 0.53 15	91.91± 1.53 18	90.02± 1.82 21	86.35± 2.03 20	87.96± 2.34 25	**98.48± 0.68 09**
Lung	#Acc ± STD # Genes	97.05 ±0.87 20	95.61 ±0.53 16	88.46 ±1.31 19	87.96 ±2.02 24	91.25 ±0.92 25	**99.45± 0.05 15**
Breast	#Acc ± STD # Genes	93.64 ±0.87 21	92.17 ±0.96 17	90.99 ±1.52 15	91.68 ±1.63 20	88.47 ±2.13 18	**98.87 ±1.10 12**
DLBCL	#Acc ± STD # Genes	95.42 ±1.52 25	97.52 ±0.36 18	94.76 ±2.03 24	96.21 ±1.61 22	90.52 ±2.83 26	**99.62± 0.08 18**
Arcene	#Acc ± STD # Genes	90.41 ±1.03 17	91.92 ±0.86 19	82.65 ±2.31 23	87.36 ±1.53 21	89.56 ±2.52 25	**96.98± 0.08 11**
Dorothea	#Acc ± STD # Genes	87.23 ±2.89 27	94.15 ±2.51 22	86.32 ±2.74 18	87.35 ±2.57 24	84.36 ±2.63 27	**98.52 ±0.84 12**

The proposed technique is also compared with some state of arts hybrid techniques such as IWSSr+ Shuffled Frog Leaping Algorithm (SFLA) and teaching learning-based algorithm gravitational search algorithm (TLB-GSA) in Table 6. The results show the proposed method outperforms the two other hybrid techniques except in the prostate and DLBCL datasets where the accuracy of TLBOGSA and IWSSr+SFLA is comparatively higher. The highest accuracy achieved in the Lung cancer dataset with the proposed method is 99.45% with 15 optimal numbers of genes. Table 7 depicts the comparison of accuracy, sensitivity, specificity, F-measure, MCC, and kappa statistic of the optimal subset of genes obtained after applying mRMR and IBCSO algorithm is classified with four different classifiers including ANN (Optimized with WOA), the comparative evaluation states that ANN (WOA) provides more promising results. The Kappa Statistics is the highest (0.974) in the ANN classifier based on the whale optimization algorithm.

Table 6. Comparison of the proposed technique with other hybrid techniques.

Dataset	Performance	Proposed method	IWSSr+SFLA	TLBOGSA
Prostate	#Acc ± STD # Genes	97.48 ± 0.68 09	95.18 ±0.58 08	**98.42 ± 0.67 07**
Lung	#Acc ± STD # Genes	**99.45± 0.05 15**	98.16 ± 0.21 12	99.10 ± 0.02 13
Breast	#Acc ± STD # Genes	**98.87±1.10 12**	90.17 ± 0.14 11	97.87 ± 0.10 13
DLBCL	#Acc ± STD # Genes	**99.34 ± 0.08 18**	98.21 ± 0.47 15	97.26± 0.86 20
Arcene	#Acc ± STD # Genes	96.98 ± 0.08 11	**97.36 ± 0.78 9**	95.84 ± 0.84 12
Dorothea	#Acc ± STD # Genes	**98.52 ±0.84 12**	92.43 ± 0.56 21	96.87 ± 0.10 14

Table 7. Average classification performance of the proposed method using four different classifiers on six datasets.

Dataset	Measures	SVM	NB	kNN	ANN(WOA)
Prostate	Acc	94.44	94.78	90.65	97.48
	Se	96.25	93.25	89.65	96.38
	Sp	91.81	91.56	91.03	95.32
	Fmes	90.32	88.98	89.57	98.10
	MCC	0.89	0.90	0.88	0.97
	Kappa	0.88	0.89	0.86	0.95
Lung	Acc	98.16	94.97	92.87	99.45
	Se	99.10	94.11	91.35	99.21
	Sp	96.60	93.36	91.80	97.12
	Fmes	95.30	93.01	91.54	98.23
	MCC	0.97	0.94	0.95	0.98
	Kappa	0.96	0.88	0.87	0.97

Dataset	Measures	SVM	NB	kNN	ANN(WOA)
Breast	Acc	87.89	86.87	82.98	98.87
	Se	89.00	85.69	80.52	95.89
	Sp	85.55	88.07	7935	93.45
	Fmes	81.20	86.95	78.99	97.56
	MCC	0.93	0.94	0.90	0.97
	Kappa	0.84	0.83	0.80	0.96
DLBCL	Acc	98.12	96.07	94.88	99.34
	Se	98.33	95.36	95.36	99.12
	Sp	95.00	94.87	93.89	98.10
	Fmes	95.26	93.65	92.67	98.20
	MCC	0.95	0.94	0.91	0.96
	Kappa	0.96	0.92	0.90	0.96
Arcene	Acc	94.00	93.21	91.32	96.98
	Se	92.21	92.36	90.55	94.89
	Sp	95.45	90.35	89.65	92.56
	Fmes	92.46	91.58	88.39	95.32
	MCC	0.90	0.90	0.92	0.95
	Kappa	0.89	0.88	0.87	0.91
Dorothea	Acc	90.37	90.72	86.38	98.52
	Se	91.31	88.36	85.75	96.45
	Sp	89.47	89.65	84.96	93.12
	Fmes	86.97	91.56	85.32	98.13
	MCC	0.89	0.93	0.89	0.97
	Kappa	0.86	0.86	0.83	0.97

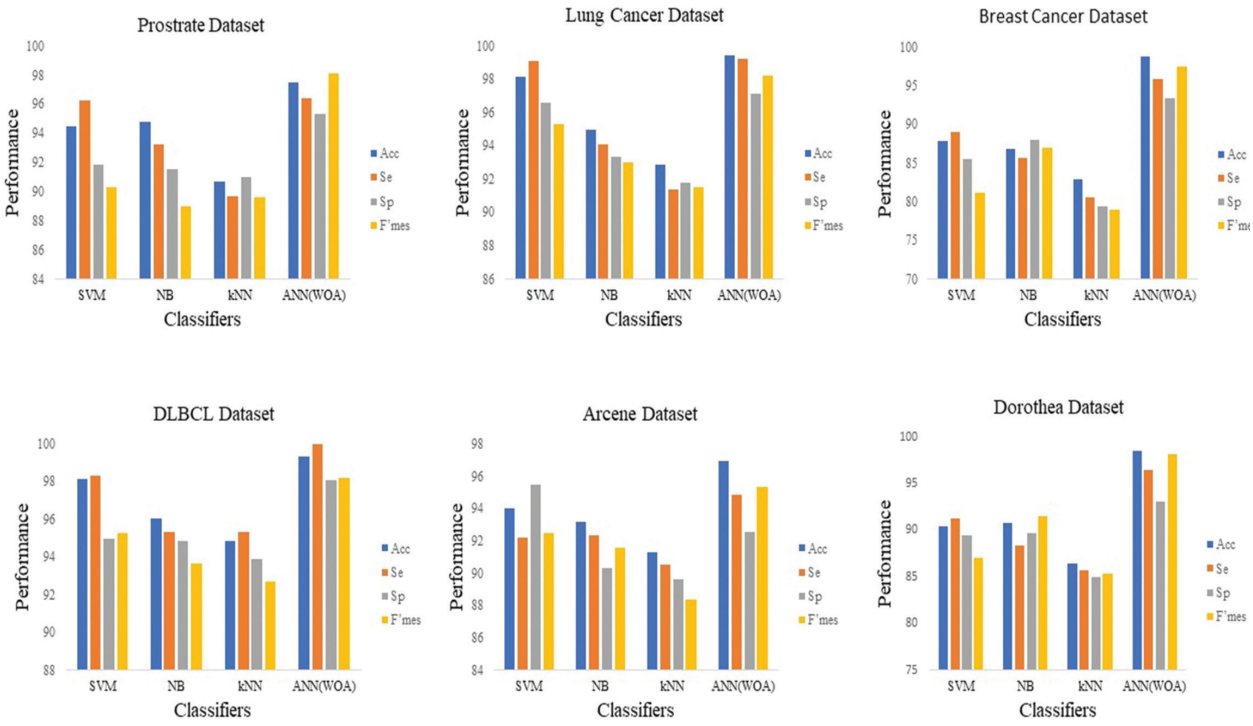

Figure 3. Performance comparison of the proposed technique using different classifiers on all the datasetsa

Fig. 3 displays the performance observation of the proposed gene selection technique with different classifiers including ANN(WOA) on all datasets. The parameters used are sensitivity (S_e), specificity (S_p), Accuracy (Acc), and F-measure (F'_{mes}).

5. CONCLUSION

Recently in the field of computational biology researchers are being attracted to the identification of marker genes related to complex diseases, especially

cancer diagnosis. However, it has been a difficult task to identify those markers due to the high dimensionality of microarray datasets. Although several existing techniques are efficient to strike out informative features from large datasets, these methods have some shortcomings such as slow convergence rate and high computational cost. To attain a good balance between exploration and exploitation and identify informative genes, a hybrid technique called IBCSOWOA is proposed to accelerate the gene selection process and improve classification accuracy. The IBCSOWOA incorporates the qualities of IBCSO and WOA techniques, the IBCSO selects the relevant feature subset and WOA optimizes the hyperparameters of ANN to improve the classification accuracy. The proposed technique is also introducing a new fitness function for identifying the informative genes. The experiments are conducted on six different biological datasets and the results illustrate that the technique outperforms other existing methods in terms of relevant gene subset selection and classification accuracy and also reduced the computational time. Out of six datasets, the proposed technique achieves more than 98% accuracy in all datasets. Therefore, it can be concluded that the proposed technique has been able to enhance the classification performance and reduce the computational time. This study also has some potential limitations such as all the datasets are microarray type and low sample size as compared to other gene datasets in the future RNA-seq datasets can be applied, as they are less noisy and more accurate. The current work can further be enhanced to incorporate deep learning techniques to improve the classification process.

Conflict of Interest

The authors of this publication declare there is no conflict of interest.

Funding Agency

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Funding for this research was covered by the author(s) of the article.

Author Contribution

All authors have contributed equally.

6. REFERENCES

[1] M. S. Lawrence, P. Stojanov, P. Polak, C. Stewart, Y. Drier, E. Helman, J. Kim, G. Getz, "Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes", Nature, Vol. 499, No. 7457, 2013, pp. 214-218.

[2] The International Cancer Genome Consortium, "International Network of Cancer Genome Projects", Nature, Vol. 464, No. 7291, 2010, pp. 993-998.

[3] L. Chin, M. L. Meyerson, "Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways", Nature, Vol. 455, No. 7216, 2008, pp. 1061-1068.

[4] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gassenbeek, J. P. Messirov, H. Coller, E. S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", Science, Vol. 286, No. 5439, 1999, pp. 531-537.

[5] G. Isabelle, W. Jason, B. Stephen, V. Vladimir, "Gene Selection for Cancer Classification Using Support Vector Machines", Machine Learning, Vol. 46, No. 1, 2002, pp. 389-422.

[6] A. Sharma, R. Rani, "KSRMF: Kernelized Similarity Based Regularized Matrix Factorization Framework for Predicting Anti-Cancer Drug Responses", Journal of Intelligent & Fuzzy Systems, Vol. 35, No. 2, 2018, pp. 1779-1790.

[7] M. Phogat, D. Kumar, "Disease Single Nucleotide Polymorphism Selection Using Hybrid Feature Selection Technique", Journal of Physics: Conference Series, Vol. 1950, No. 1, 2021, p. 012079.

[8] A. K. Shukla, P. Singh, M. Vardhan, "An Adaptive Inertia Weight Teaching-Learning-Based Optimization Algorithm and Its Applications", Applied Mathematical Modelling, Vol. 77, 2020, pp. 309-326.

[9] F. S Gharehchopogh, H. Gholizadeh, "A Comprehensive Survey: Whale Optimization Algorithm and Its Applications", Swarm and Evolutionary Computation, Vol. 48, 2019, pp. 1-24.

[10] E. Zorarpaci, S. A. Ozel, "A Hybrid Approach of Differential Evolution and Artificial Bee Colony for Feature Selection", Expert Systems with Applications, Vol. 62, 2016, pp. 91-103.

[11] M. Phogat, D. Kumar, "Classification of Complex Diseases Using an Improved Binary Cuckoo Search and Conditional Mutual Information Maximization", Computación y Sistemas, Vol. 24, No. 3, 2020.

[12] E. Pashaei, N. Aydin, "Gene Selection Using Hybrid Binary Black Hole Algorithm and Modified Binary Particle Swarm Optimization", Genomics, Vol. 111, No. 4, 2019, pp. 669-686.

[13] A. K. Shukla, P. Singh, M. Vardhan, "Gene Selection for Cancer Types Classification Using Novel Hybrid Metaheuristics Approach", Swarm and Evolutionary Computation, Vol. 54, 2020, p. 100661.

[14] S. Dwivedi, M. Vardhan, S. Tripathi, A. K. Shukla "Implementation of Adaptive Scheme in Evolutionary Technique for Anomaly-Based Intrusion Detection", Evolutionary Intelligence, Vol. 13, No. 1, 2020, pp. 103-117.

[15] H. M. Alshamlan, G. H. Badr, Y. A. Alohali, "Genetic Bee Colony (GBC) Algorithm: A New Gene Selection Method for Microarray Cancer Classification", Computational Biology and Chemistry, Vol. 56, 2015, pp. 49-60.

[16] W. M. Shaban, "Insight into Breast Cancer Detection: New Hybrid Feature Selection Method", Neural Computing & Applications, Vol. 35, 2023, pp. 6831-6853.

[17] N. Kumar, D. Kumar, "An Improved Grey Wolf Optimization-Based Learning of Artificial Neural Network for Medical Data Classification", Journal of Information and Communication Technology, Vol. 20, No. 2, 2021, pp. 213-248.

[18] A. Sharma, R. Rani, "C-HMOSHSSA: Gene Selection for Cancer Classification Using Multi-Objective Meta-Heuristic and Machine Learning Methods", Computer Methods and Programs in Biomedicine, Vol. 178, 2019, pp. 219-235.

[19] K. Adem, S. Kiliçarslan, O. Comert, "Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification", Expert Systems with Applications, Vol. 115, 2019, pp. 557-564.

[20] A. Sampathkumar, R. Rastogi, S. Arukonda, A. Shankar, S. Kautish, M. Sivaram, "An efficient hybrid methodology for detection of cancer-causing gene using CSC for microarray data, Journal of Ambient Intelligence and Humanized Computing, Vol. 11, 2020, pp. 4743-4751.

[21] S. Akila, S.A. Christe, "A wrapper-based binary bat algorithm with greedy crossover for attribute selection", Expert System with Applications, Vol. 187, 2022, p. 115828.

[22] M. A. Al-Betar, O. A. Alomari, S. M. Abu-Romman, "A TRIZ-inspired bat algorithm for gene selection in cancer classification", Genomics, Vol. 112, 2020, pp. 114-126.

[23] M. Hamim, I. El Moudden, M. D. Pant, H. Moutachaouik, M. Hain, "A Hybrid Gene Selection Strategy Based on Fisher and Ant Colony Optimization Algorithm for Breast Cancer Classification", International Journal of Online and Biomedical Engineering, Vol. 17, 2021, p. 148.

[24] G. G. Afif, Adiwijaya, W. Astuti, "Cancer Detection Based on Microarray Data Classification Using FLNN and Hybrid Feature Selection", Journal RESTI, Vol. 5, 2021, pp. 794-801.

[25] A. Tahmouresi, E. Rashedi, M. M. Yaghoobi, M. Rezaei, "Gene Selection Using Pyramid Gravitational Search Algorithm", PLoS ONE, Vol. 17, 2022, p. e0265351.

[26] R. Cheng, Y. Jin, "A Competitive Swarm Optimizer for Large Scale Optimization", IEEE Transactions on Cybernetics, Vol. 45, No. 2, 2015, pp. 191-204.

[27] S. Mirjalili, A. Lewis, "The Whale Optimization Algorithm", Advances in Engineering Software, Vol. 95, 2016, pp. 51-67.

[28] W. S. McCulloch, W. Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity", The Bulletin of Mathematical Biophysics, Vol. 5, No. 4, 1943, pp. 115-33.

[29] J. Khan, J. S. Wei, L. H Saal, M. Ladanyi, F. Wastermann, P. S. Meltzer, "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks", Nature Medicine, Vol. 7, No. 6, 2001, pp. 673-679.

[30] H. C. Wu, X. G. Wei, S. C. Chan, "Novel Consensus Gene Selection Criteria for Distributed GPU Partial Least Squares-Based Gene Microarray Analysis in Diffused Large B Cell Lymphoma (DLBCL) and Related Findings", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 15, No. 6, 2018, pp. 2039-2052.

[31] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, F. Herrera, "A Review of Microarray Datasets and Applied Feature Selection Methods", Information Sciences, Vol. 282, 2014, pp. 111-135.

[32] G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, R. Bueno, "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma", Cancer Research, Vol. 62, No. 17, 2002, pp. 4963-4967.

[33] M. A. Shipp, K. N Ross, P. Tamayo, A. P Weng, J. L Kutok, M. Angelo, M. Reich, T. R. Golub, "Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning", Nature Medicine, Vol. 8, No. 1, 2002, pp. 68-74.